Effective Approach for Disambiguating Chinese Polyphonic Ambiguity

Feng-Long Huang
Department of Computer Science and Information Engineering
National United University
No. 1, Lienda, Miaoli, Taiwan, 36003
flhuang@nuu.edu.tw

***Abstract:*-One of the difficult tasks on Natural Language Processing (NLP) is to resolve the sense ambiguity of characters or words on text, such as polyphones, homonymy, and homograph. The paper addresses the ambiguity issue of Chinese character polyphones and disambiguity approach for such issues. Three methods, dictionary matching, language models and voting scheme, are used to disambiguate the prediction of polyphones. Compared with the well-known MS Word 2007 and language models (LMs), our approach is superior to these two methods for the issue. The final precision rate is enhanced up to 92.75%. Based on the proposed approaches, we have constructed the e-learning system in which several related functions of Chinese transliteration are integrated.**

***Keywords:*-Natural Language Processing, Sense Disambiguity, Language Model, Voting Scheme**,

I. INTRODUCTION

In recent years, natural language processing (NLP) has been studied and discussed on many fields, such as machine translation, speech processing, lexical analysis, information retrieval, spelling prediction, hand-writing recognition, and so on [1][2]. In the computational models, syntax models parsing, word segmentation and generation of statistical language models have been the focus tasks.

In general, no matter what kinds of natural languages, there will be always a phenomenon of ambiguity among characters or words in text, such as polyphone, homonymy, homograph, and the combination of them. It is of necessary to accomplish most natural language processing applications. One of the difficult tasks on NLP is to resolve the word's sense ambiguity. It is so-called word sense dsiambiguity (WSD) [3, 4].

Disambiguating the sense ambiguity can alleviate the problems in NLP. The paper address the dictionary matching, statistical *N*-gram language model (LMs) and voting scheme, which includes two methods: preference and winner-take-all scoring, to retrieve Chinese lexical knowledge, employed to process WSD on Chinese polyphonic characters. There are near 5700 frequent unique characters and among them more than 1300 characters have more than 2 different pronunciations, they are called polyphonic characters. The problem predicting correct polyphonic categories can be regarded as the issue of WSD.

The paper is organized as following: the related works on WSD are presented in Section 2. Three methods will first be described in Section 3 and experimental results are shown and then analyzed furthermore in Section 4. Conclusions and future works are listed in last section.

II. RELATED WORKS

Resolving automatically the word sense ambiguity can enhance the language understanding, which will used on several fields, such as information retrieval, document category, grammar analysis, speech processing and text preprocessing, and so on. In the past decades, ambiguity issues are always considered as AI-complete, that is, a problem which can be solved only by first resolving all the difficult problems in artificial intelligence (AI), such as the representation of common sense and encyclopedic knowledge. Sense disambiguation is required for correct phonetization of words in speech synthesis [13], and also for word segmentation and homophone discrimination in speech recognition.

It is essential for language understanding applications suchas message understanding, man-machine communication, etc. WSD can be applied into many fields of natural language processing [10], such as machine translation, information retrieval (IR), speech processing and text processing.

The approaches on WSD are categorized as follows:

A. Machine-Readable Dictionaries (MRD):

Relying on the word information in dictionary for sense ambiguity, such as WordNet or Academia Sinica Chinese Electronic Dictionary (ASCED) [17].

B. Computational Lexicons:

Employing the lexical information in thesaurus, such as the well-known WordNet [11, 14], which contains the lexical clues of characters and lattice among related characters.

C. Corpus-based methods

Depending on the statistical results in corpus, such as term's occurrences, part-of-speech (POS) and location of characters and words [12, 15].

D. Neural Networks:

The approach is based on the concept codes of thesaurus or features of lexical words [16, 17].

There are many works addressing WSD and several methods have been proposed so far. Because of the unique features of Chinese language-Chinese word segmentation, more than two different features will be employed to achieve higher prediction for WSD issues. Therefore, two methods will be arranged furthermore.

DOI: 01.IJSIP.02.01.211

III. Description of Proposed Methods

In this paper, several methods are first proposed to disambiguate the sense category of Chinese polyphones; dictionary matching, n-gram language models and voting scheme. In the following, each will be explained in details.

A. Dictionary Matching

In order to predict correctly the sense category of polyphones, dictionary matching will be exploited for the ambiguity issue. Within a Chinese sentence, the location *p* of polyphonic character w_p is set as the centre, we extract the right and left substring based on the centre *p*. Two substrings are denoted as CH_L and CH_R, as shown in Fig. 1. In a window size, all possible substrings in CH_L and CH_R will be segmented and then match the lexicons in dictionary.

w_1, w_2 (CH_L) $\mathbf{w_p}$

$w_1, w_2 \ldots\ldots$ $\mathbf{w_p}$ $w_{p+1}, w_{p+2} \ldots\ldots$ w_n

$\mathbf{w_p}$ w_{p+1}, w_{p+2} (CH_R) w_n

Fig. 1: A sentence with target polyphonic character $\mathbf{w_p}$. will divided into two substrings.

If the words are existed on both substrings, then we can decide the pronunciation of polyphone based on the priority of longest word and highest frequency of word; length of word first and then frequency of word secondly. In the paper, window size=6 Chinese characters; that means LEN(CH_L)= LEN(CH_R)=6。

The Chinese dictionary is available and contains near 130K Chinese words. Each Chinese word may be composed from 2 to 12 Chinese characters. All the words in dictionary contain its frequency, part-of-speech (POS), transliteration[1]; in which correctly pronunciation for polyphonic character in the word may be decided.

The algorithm of dictionary matching is described as follows:

step 1. Read in the sentence and find the location *p* of polyphone target w_p.

step 2. Based on the of w_p, all the possible substring of CH_L and CH_R within window (size=6) will be segmented and extracted, then compared with lexicons in Chinese dictionary.

step 3. If any Chinese word can be found on both substring goto step 4,
else
goto step 5.

step 4. Decide the sense category of pronunciation for polyphone based on the priority scheme of longest word and highest frequency of word. Then the process ends.

step 5. The pronunciation of polyphone w_p will be predicted by methods in the following phase.

B. Language Models - LMs

In recent years, the statistical language models have been adopted in NLP. Supoosed that $W=w_1,w_2,w_3,\ldots w_n$, where w_i and n denote the the i^{th} Chinese character and number of characters in sentence $(0 \le i \le n)$。

$P(W)=P(w_1, w_2 \ldots\ldots, w_n)$, *//using c*hain rules.

$$P(w_1^n)= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1})$$
$$=\prod_{k=1}^{n} P\left(w_k \middle| w_1^{k-1}\right) \tag{1}$$

where w_1^{k-1} denotes string $w_1,w_2,w_3,\ldots w_{k-1}$.

In Eq(1), the probability $P\left(w_k \middle| w_1^{k-1}\right)$ can be calculated, starting at w_1, by using $w_1,w_2,w_3\ldots w_{k-1}$ substring to predict the occurrence probability of w_k. In case of longer string, it is necessary for large amount of corpus to train the language model with better performance. It will lead to spending much labor and time extensive.

In general, unigram, bigram and trigram (3<=*N*) [5][6] are generated. *N*-gram model calculates probability $P(\ .\ )$ of N^{th} events by the preceding *N*-1 events, rather than string $w_1,w_2,w_3\ldots w_{N-1}$.

In short, *N*-gram is so-called N-1)th-order Markov model, which calculate conditional probability of successive events: calculate the probability of N^{th} event while preceding (*N*-1) event occurs. Basically, *N*-gram Language Model is expressed as follows:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1}) \tag{2}$$

N=1, unigram or zero-order markov model.
N=2, bigram or first-order markov model.
N=3, trigram or second-order markov model.

In Eq(2), the relative frequency will be used for calculating the $P(\ .\ )$:

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}, \tag{3}$$

where $C(w)$ denotes the count of event *w* occurring in training corpus.

In Eq(3), the obtained probability $P(\ .\ )$ is called Maximum Likelihood Estimation (MLE). While predicting the pronunciation category of polyphones, we can predict based on the probability on each category t $(1 \le t \le T)$, T denotes the number of categories of polyphone. The category with maximum probability $P_{max}(W)$ with respect to the sentence *W* will be the target and then the correct pronunciation of polyphone can be decided.

C. Voting Scheme

In contrast to the *N*-gram models above, we proposed voting scheme with similar concept for use to select in human being society. Basically, we vote for one candidate and the candidates with maximum votes will be the winner. In real world, maybe more than one candidate will win the section game while disambiguation process only one

[1] Zhuyin Fuhau (注音符號) can be found in the dictionary.

DOI: 01.IJSIP.02.01.211

category of polyphone will be the final target with respect to the pronunciation.

The voting scheme can be described as follows: each token in sentence play the voter for vote for favorite candidate based on the probability calculated by the lexical features of tokens. The total score $S(W)$ accumulated from all voters for each category will be obtained, and the candidate category with highest score is the final winner. In the paper, there are two voting methods:

1) Winner-Take-All:

In the voting method, the probability is calculated as follows:

$$P(w_i) = \frac{C(w_i,t)}{C(w_i)} \tag{4}$$

where $C(w_i)$ denotes the occurrences of w_i in training corpus, and $C(w_i, t)$ denotes the occurrences of w_i for sense category t in training corpus.

In Eq(4) above, $P(w_i)$ is regarded as the probability of w_i on category t. In winner take all scoring, the category with maximum probability will win the ticket. On the other hand, it win one ticket (1 score) while all other categories can't be assigned any ticket (0 score). Therefore, each voter has just one ticket for voting. The winner-take-all scoring for tolen w_i can be defined as follows:

$$P_t(w_i) = \begin{cases} 1 & \text{if } P_t(w_i) = \max.\ \text{among all categories } T \\ 0 & \text{all other categories} \end{cases} \tag{5}$$

According to Eq(5), the total score for each categories can be accumulated for all tokens in sentence:

$$S(W) = P(w_1)+P(w_2)+P(w_3)+\ldots\ldots+P(w_n)$$
$$=\sum_{k=1}^{n} P(w_k) \tag{6}$$

2) Preference Scoring:

Another voting method is called as preference. For a token in sentence, the summation of the probability for all the categories of a polyphone character will be equal to 1. Let us show an Chinese's example (E1) for two voting methods. Note that sentence (E1') is the translation for example (E1). As presented in Table 1, the polyphone character 卷 has three different pronunciations, 1. ㄐㄩㄢˋ, 2. ㄐㄩㄢˇ and 3. ㄑㄩㄢˊ. Supposed that the occurrence of token 白卷 (blank examination) in these phonetic categories are 26, 11 and 3, total occurrence is 40. Therefore, the score for each category by two scoring methods can be calculated.

教育社會方面都繳了白卷 (E1)

Government handed over a blank examination paper in education and society. (E1')

Table 1: example for two scoring scheme of voting.

category	count	preference	w-t-all
1 ㄐㄩㄢ 4	26	26/40=0.65	40/40=1
2 ㄐㄩㄢ 3	11	11/40=0.275	0/40=0
3 ㄑㄩㄢ 2	3	3/40=0.075	0/40=0
Total $\sum C()$	40	1 score	1 score

ps. w-t-all denotes winner-take-all scoring

D. Unknown events-Zero Count Issue

In certain cases, $C(\bullet)$ of a novel (unknown word), which don't occur in the training corpus, may be zero because of the limited training data and infinite language. It is always hard for us to collect sufficient datum. The potential issue of MLE is the probability for unseen events is exactly zero. This is so-called the zero-count problem and will degrade the performance of system.

It is obvious that zero count will lead to the zero probability of $P(\bullet)$ in Eqs(2), (3) and (4). There are many smoothing works in [7, 8, 9]. The paper adopted the additive discounting for calculating P^* as follows:

(7)

where δ denotes a small value (δ<=0.5); which will be added into all the known and unknown events. The smoothing method will alleviate the zero count issue in language model.

E. Classifier-Predicting the Categories

Supposed that polyphone has T categories, $1 \leq t \leq T$, how can we predict the correct target $\hat{t}$? As shown in Eq(8), the category with maximum probability or score will be the most possible target:

$$\hat{t} = argmax_t P_t(W), \quad \text{or}$$
$$\hat{t} = argmax_t S_t(W), \tag{8}$$

where $P_t(W)$ is the probability of W in category t, which can be obtained from Eq(1) for LMs and $S_t(W)$ is the total score based on the voting scheme from Eq(6).

IV. Experiment Results

In the paper, 10 Chinese polyphones are selected randomly from more than 1300 polyphones in Chinese. All the promising pronunciations of these selected polyphones are list in Table 2; one polyphone "著" has 5 categories, 3 polyphone have 2 categories.

A. Dictionary and Corpus

Academic Sinica Chinese Electronic dictionary, ASCED) contains more than 130K Chinese words, composing of 2 to 11 characters. The word in ASCED is with Part-of-speech (POS), frequency and pronunciation for each character.

The experimental data are collected from the corpus of ASBC (Academia Sinica Balanced Corpus) and web news of China Times. The sentences with one of 10 polyphones are collected randomly. There are totally 9070 sentences,

DOI: 01.IJSIP.02.01.211

which are divided into two parts: 8030 (88.5%) and 1040 (11.5%) sentences for training and outside testing, respectively.

B. *Experiment Results*

Three LMs models are generated: unigram, bigram and trigram. Precision Rate (PR) can be defined as:

$$PR = \frac{\text{NO. of correct prediction}}{\text{total number of sentence}} \qquad (9)$$

Method 1: Dictionary Matching

There are 69 sentences processed by the word matching phase and 7 sentences are wrongly predicted. The average PR achieves 89.86%.

In the followings, several examples are presented and explained the matching phase of dictionary matching:

我們回頭看看中國人的歷史。 (E2)

We look back the history of Chinese. (E2')

Based on the matching algorithm, two substring CH_L and CH_R of polyphone target(w_p =**中**) for sentence (E2);

CH_L =”們回頭看看中”,

CH_R=”中國人的歷史”.

Upon the word segmentation, the Chinese word and pronunciation are as follows:

CH_L			CH_R		
看中	83	ㄓㄨㄥ 4	中國	3542	ㄓㄨㄥ
			中國人	487	ㄓㄨㄥ

According the priority of length of word first,中國人 (Chinese people) will decide the pronunciation of 中 as ㄓㄨㄥ.

看中文再用廣東話來發音。 (E3)

Read the Chinese and then pronounce in Canton. (E3')

Chinese words in CH_L			Chinese words in CH_R		
看中	83	ㄓㄨㄥ 4	中文	343	ㄓㄨㄥ

峰迴路轉再看中國方面. (E4)

The path winds along mountain ridges, then watch the reflection of China. (E4')

Chinese words in CH_L			Chinese words in CH_R		
看中	83	ㄓㄨㄥˋ	中國	3542	ㄓㄨㄥ

中央研究院未來的展望。 (E5)

The future forecast of Academic Sinica of Chinese. (E5')

Chinese words in CH_L	CH_R		
NULL	中央	2979	ㄓㄨㄥ
	中央研究院	50	ㄓㄨㄥ

In example (E5), only CH_R contains the segmented words. On the other hand, there are no any word in CH_L

Method 2: Language Model (LMs)

The experiment results of three models unigram, bigram, trigram are listed in Table 3. Bigram LMs achieves 92.58%, which is highest rate among three models.

Method 3: Voting Scheme

1)Winner take all: Three models; unitoken, betoken and tritoken are generated. As shown in Table 4. Bitoken achieves highest PR of 90.17%.

2)Preference**:** Three models; unitoken, bitoken and tritoken are generated. As shown in Table 5. Bitoken preference scoring can achieves highest PR of 92.72% in average.

C. *Word 2007 precision rate*

MS Office is a famous and well-known editing package around world. In our experiments, MS Word 2007 is used to process the transcription on same testing sentences. PR achieves 89.8% in average, as shown in Table 6.

D. *Results Analysis*

In the paper, voting scheme of preference and winner-take-all scoring, and statistical language Model have been proposed and employed to resolve the issue of polyphone ambiguity. We compare these methods with MS Word 2007. Preference bitoken scheme achieves highest PR among these models and achieves 92.72%. It is apparent that all our proposed methods are superior to MS Word 2007.

In the following, two examples are shown for correct and wrong prediction by Word 2007.

ㄐㄧㄠˋ ㄩˋ ㄕㄜˋ ㄏㄨㄟˋ ㄈㄤ ㄇㄧㄢˋ ㄉㄡ ㄐㄧㄠˇ ˙˙ ㄅㄞˊ ㄐㄩㄢˋ
教 育 社 會 方 面 都 繳 了 白 卷

Government handed over a blank examination paper in education and society. (correct prediction)

ㄅㄤ ㄖㄨㄛˋ ㄨˊ ㄖㄣˊ ㄅㄢ ㄗˋ ㄧㄢˊ ㄗˋ ㄩˇ
傍 若 無 人 般 自 言 自 語

Talking to oneself as if nobody is around.(wrong prediction)

We have constructed an intelligent e-learning system [18] based on the unify approach proposed in the paper. The system provides the function of Chinese Synthesized speech and display sereral useful lexical information, such as transliteration, Zhuyin and 2 pinyins for learning Chinese.

All the functions such as Chinese polyphones prediction addressed in the paper, transliteration and transcription described above are integrated together in the e-learning website to provide online searching and translation through

DOI: 01.IJSIP.02.01.211

Internet. If the predicted category is wrong, user may feedback the right category of polyphone to online gradually adapt the system's prediction for Chinese polyphones.

V. Conclusion

In the paper, we used several methods to address the issue of ambiguity of Chinese polyphones. First, three methods are employed to predict the category of polyphone: dictionary matching, language models and voting scheme; the last method has two different scoring schemes: winner-take-all and preference scoring. Furthermore we propose the effective unify approaches, which unify the several methods and then adopt better alternatives triggered based on a threshold, to improve the prediction.

Our approach outperforms MS Word 2007 and statistical language models, and the best result of final outside testing achieves 92.72%. The proposed approach can be applied to related issues on other language.

Based on the proposed unify approach, we have constructed the e-learning system in which several related functions of Chinese text transliteration are integrated to provide on-line searching and translation through Internet. In future, several related issues should be studied furthermore:

1. Collecting more corpus and extend the proposed methods to other Chinese polyphones.
2. More lexical features, such as location and semantic information, used to enhance the precision rate of prediction.
3. Improving the smoothing techniques for unknown words.
4. Bilingual translation for English and Chinese.

Acknowledgement

The paper is supported partially under the Project of NCS, Taiwan.

Reference

[1] Yan Wu, Xiukun Li and Caesar Lun, 2006, A Structural Based Approach to Cantonese-English Machine Translation, Computational Linguistics and Chinese Language Processing, Vol. 11, No. 2, June 2006, pp. 137-158.

[2] Brian D. Davison, Marc Najork, Tim Converse, 2006, SIGIR Workshop Report, Vol. 40 No. 2.

[3] Oliveira, F.; Wong, F.; Li, Y.-P., 2005, Machine Learning and Cybernetics, Proceedings of 2005 International Conference on Volume 6, Issue , 18-21 Aug. 2005 Vol. 6, An unsupervised & statistical word sense tagging using bilingual sources, Page(s): 3749 - 3754

[4] Agirre E., Edmonds P., 2006, Word Sense Disambiguation Algorithms and Applications, Springer.

[5] Jurafsky D. and Martin J. H., 2000, Speech and Language Processing, Prentice Hall.

[6] Jui-Feng Yeh, Chung-Hsien Wu and Mao-Zhu Yang, 2006, Stochastic Discourse Modeling in Spoken Dialogue Systems Using Semantic Dependency Graphs, Proceedings of the COLING/ACL 2006 Main Conference Poster *Sessions*, pages 937–944.

[7] Standley F. Chen and Ronald Rosenfeld, Jan. 2000, A Survey of Smoothing Techniques, for ME Models, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, pp. 37-50.

[8] Church K. W. and Gale W. A., 1991, A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilies of English Bigrams, Computer Speech and Language, Vol. 5, pp 19-54.

[9] Chen Standy F. and Goodman Joshua, 1999, An Empirical study of smoothing Techniques for Language Modeling, Computer Speech and Language, Vol. 13, pp. 359-394.

[10] Nancy Ide and Jean Véronis, 1998, Word Sense Disambiguation, The state of the art Computational Linguistics Vol. 24, NO. 1, pp. 1-41.

[11] Miller, George A.; Beckwith, Richard T. Fellbaum, Christiane D.; Gross, Derek; and Miller, Katherine J. (1990). WordNet: A non-line lexical database. International Journal of Lexicography, **3**(4), 235-244.

[12] Church, Kenneth W. and Mercer, Robert L.(1993). Introduction to the Special Issue on Computational Linguistics using Large Corpora. Computational Linguistics, **19**(1), 1-24.

[13]Yarowsky D., Homograph disambiguation in speech synthesis. In J. van Santen, R. Sproat, J. Olive and J. Hirschberg, Progess in Speech Synthesis, Springer-Verlag, 1997, pp. 159–175.

[14] Feng-Long Huang, Shu-Yu Ke and Qiong-Wen Fan, 2008, Predicting Effectively the Pronunciation of Chinese Polyphones by Extracting the Lexical Information, *Advances in Computer and Information Sciences and Engineering*, Springer Science, pp. 159–165.

[15]Francisco Joao Pinto, Antonio Farina Martinez, Carme Fernandez Perez-Sanjulian, 2008, Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet, International Journal of Computer Applications in Technology Vol. 33, No. 4, pp. 271 – 279.

[16]You-Jin Chung et.al., 2002, Word sense disambiguation in a Korean-to-Japanese MT system using neural networks, International Conference On Computational Linguistics archive COLING-02 on Machine translation in Asia – Vol. 16, pp.1-7.

[17] Jean Veronis and Nancy M. Ide, 1990, Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, International Conference On Computational Linguistics Proceedings of the 13th conference on Computational linguistics – Vol. 2, pp. 389 -394.

[18] http://203.64.183.226/public2/word1.php

DOI: 01.IJSIP.02.01.211

Table 2**:** 10 Chinese polyphonic characters; its category and meanings.

target	Zhuyin Fuhau	Chinese word	hanyu pinyin	English
中	ㄓㄨㄥ	中心	zhong xin	center
	ㄓㄨㄥˋ	中毒	zhong du	poison
乘	ㄔㄥˊ	乘法	cheng fa	multiplication
	ㄕㄥˋ	大乘	da sheng	
乾	ㄍㄢ	乾淨	gan jing	clean
	ㄑㄧㄢˊ	乾坤	qian kun	the universe
了	ㄌㄜ˙	為了	wei le	in order to
	ㄌㄧㄠˇ	了解	liao jie	understand
傍	ㄆㄤˊ	傍邊	pang bian	beside
	ㄅㄤ	傍晚	bang wan	nightfall
	ㄅㄤˋ	依山傍水	yi shan bang shui	near the mountain and by the river
作	ㄗㄨㄛˋ	工作	gong zuo	work
	ㄗㄨㄛ	作揖	zuo yi	
	ㄗㄨㄛˊ	作興	zuo xing	
著	ㄓㄜ˙	忙著	mang zhe	busy
	ㄓㄠ	著急	zhao ji	anxious
	ㄓㄠˊ	著想	zhao xiang	to bear in mind the interest of
	ㄓㄨˋ	著名	zhu	famous
	ㄓㄨㄛˊ	執著	zhuo	inflexible
卷	ㄐㄩㄢˋ	考卷	kao juan	a test paper
	ㄐㄩㄢˇ	卷髮	Juan fa	curly hair
	ㄑㄩㄢˊ	卷曲	quan qu	curl
咽	ㄧㄢ	咽喉	yan hou	the throat
	ㄧㄢˋ	吞咽	tun yan	swallow
	ㄧㄝˋ	哽咽	geng ye	to choke
從	ㄘㄨㄥˊ	從事	cong shi	to devote oneself
	ㄗㄨㄥˋ	僕從	pu zong	servant
	ㄘㄨㄥ	從容	cong rong	calm; unhurried
	ㄗㄨㄥ	從橫	zong heng	in length and breadth

Table 3：PR of outside testing on Language Model.

token	**中**	**乘**	**乾**	**了**	**傍**	**作**	**著**	**卷**	**咽**	**從**	**avg.**
unigram	95.88	86.84	92.31	70.21	85.71	96.23	75.32	100	98	91.67	89.98
bigram	96.75	84.21	96.15	85.11	92.86	94.34	81.17	96.30	100	93.52	**92.58***
trigram	80.04	57.89	61.54	58.51	78.57	52.83	60.39	62.96	88	71.30	70.50

ps: * denotes the best PR among three *n*-gram models.

Table 4**:** PR of outside testing on Winner-take-all scoring.

token	**中**	**乘**	**乾**	**了**	**傍**	**作**	**著**	**卷**	**咽**	**從**	**avg.**
unitoken	96.96	84.21	80.77	57.45	71.43	94.34	58.44	85.19	84	87.04	84.69
bitoken	96.75	86.84	96.15	79.79	85.71	92.45	68.83	100	98	93.52	**90.17***
tritoken	79.83	60.53	61.54	60.64	78.57	52.83	59.74	66.67	88	71.3	70.69

ps: * denotes the best PR among three *n*-gram models.

DOI: 01.IJSIP.02.01.211

Table 4: PR of outside testing on Preference scoring.

token	中	乘	乾	了	傍	作	著	卷	咽	從	avg.
unitoken	96.96	84.21	80.77	70.21	71.43	94.34	70.13	85.19	88	87.96	87.76
bitoken.	96.75	86.84	96.15	87.23	85.71	93.40	81.17	100	98	93.52	**92.72***
tritoken.	80.04	60.53	61.54	60.64	78.57	52.83	59.74	66.67	88	71.30	70.78

ps: * denotes the best PR among three *n*-gram models.

Table **6:** PR of Word 2007 on same testing sentences.

token	中	乘	乾	了	傍	作	著	卷	咽	從	avg.
word 2007	93.37	76.47	76.67	83.65	78.57	93.70	78.33	82.76	100	91.51	**89.80**

DOI: 01.IJSIP.02.01.211